

## Claims

1. A method for gene mapping from chromosome and phenotype data, which utilizes linkage disequilibrium between genetic markers  $m_i$ , which are polymorphic nucleic acid or protein sequences or strings of single-nucleotide polymorphisms deriving from a chromosomal region, wherein
  - i) all marker patterns  $P$  that satisfy a pattern evaluation function  $e(P)$  are searched from the data, wherein
    - a. the marker patterns are expressions involving the genetic markers and their alleles and zero or more of the following: individual covariates, environmental variables and auxiliary phenotypes; and
    - b. the pattern evaluation function  $e(P)$  involves some statistical measure of the association between the marker pattern  $P$  and the phenotype being studied,
  - ii) each marker  $m_i$  of the data is scored by a marker score  $s(m_i)$ , which is a function of the set  $S_i$  defined as the set of marker patterns overlapping the marker  $m_i$  and satisfying the pattern evaluation function  $e$  as defined in step (i), and
  - iii) the location of the gene is predicted as a function of the scores  $s(m_i)$  of all the markers  $m_i$  in the data and is based on maximizing the score if the scoring function is designed to give higher scores closer to the gene, and on minimizing the score if the scoring function is designed to give lower scores closer to the gene, as is the case for instance when the scores  $s(m_i)$  are marker-wise p values.
2. A method of claim 1, wherein the chromosome data consists of either haplotypes or genotypes.
3. A method of claim 1, wherein the haplotypes and genotypes referred to in the marker patterns contain flexible regions such as gaps or disjunctions.
4. A method of claim 1, wherein the marker patterns  $P$  are searched by the following algorithm:

### Input

- set  $U$  of marker patterns

- evaluation function  $e(P)$  for patterns  $P$  in  $U$
- (generalization) relation  $<$  for patterns in  $U$
- where the function  $e$  and the relation  $<$  are such that if  $e(P)$  is true and  $P' < P$ , then  $e(P')$  is also true

## 5 Output

- set  $S = \{P \in U \mid e(P) \text{ is true}\}$  of patterns

### Method

```

1.  $S := \{\}$ 
2. // Initialize the set of evaluated patterns:
10 3.  $E := \{\}$ 
4. // Start with the most general patterns:
5.  $Gen := \{P \text{ in } U \mid \text{there is no } P' \text{ in } U, P' \neq P, \text{ such that } P' < P\}$ 
6. // Recursively evaluate patterns in a depth first order:
7. foreach  $P \in Gen$  { evaluatePatterns( $P$ ) }
15 8. end;

9. procedure evaluatePatterns( $P$ ) {
10.   insert  $P$  into the set  $E$ 
11.   if  $e(P) = \text{true}$  then {
20 12. insert  $P$  into set  $S$ 
13. // Find all specializations of  $P$  that have not been tested yet, and
14. // evaluate them recursively:
15.  $Spec := \{P' \text{ in } U-E \mid P < P', P' \neq P, \text{ and there is no } P'' \text{ in } U-E, P'' \neq P$ 
16.        $\text{and } P'' \neq P', \text{ with } P < P'' < P'\}$ ;
25 17. foreach  $P' \text{ in } Spec$  { evaluatePatterns( $P'$ ); }
18.   }
19. }
```

5. A method of claim 1, wherein the marker patterns  $P$  are searched by the following algorithm:

## 30 Input

- set  $U$  of marker patterns
- evaluation function  $e(P)$  for patterns  $P$  in  $U$
- frequency threshold  $x$

## Output

- set  $S = \{P \text{ in } U \mid e(P) \text{ and } ae(P) \text{ is true}\}$  of patterns, where  $ae(P)$  is true if and only if the frequency of pattern  $P$  exceeds a given threshold  $x$

## Method

```

5  20.  $S := \{\}$ 
   21. // Initialize the set of evaluated patterns:
   22.  $E := \{\}$ 
   23. // Start with the most general patterns:
   24.  $Gen := \{P \text{ in } U \mid \text{there is no } P' \text{ in } U, P' \neq P, \text{ such that } P \rightarrow P'\}$ 
10 25. // Recursively evaluate patterns in a depth first order:
   26. foreach  $P$  in  $Gen$  { evaluatePatterns( $P$ ) }
   27. end

   28. procedure evaluatePatterns( $P$ ) {
15 29.   insert  $P$  into the set  $E$ 
   30.   if  $ae(P) = \text{true}$  then {
   31.     if  $e(P) = \text{true}$  then insert  $P$  into set  $S$ 
   32.     // Find all specializations of  $P$  that have not been tested yet, and evaluate
   33.     // them recursively:
20 34.      $Spec := \{P' \text{ in } U-E \mid P' \rightarrow P, P' \neq P, \text{ and there is no } P'' \text{ in } U-E, P'' \neq P$ 
   35.          $\text{and } P'' \neq P', \text{ with } P' \rightarrow P'' \text{ and } P'' \rightarrow P\}$ 
   36.     foreach  $P'$  in  $Spec$  { evaluatePatterns( $P'$ ) }
   37.   }
   38. }

25 6. A method of claim 1, wherein the marker patterns  $P$  are searched by the fol-
   lowing algorithm:

```

## Input

- marker map  $M = (m_1, \dots, m_k)$
- phenotype vector  $Y = (Y_1, \dots, Y_n)$
- 30 • haplotype matrix  $H$  of size  $n * k$
- association threshold  $x$  for chi-squared test
- maximum pattern length  $l$
- maximum number of gaps  $g$
- maximum gap size  $s$

## Output

- set  $S = \{P \text{ in } U \mid e(P) \text{ is true}\}$  of patterns,
- where  $U$  consists of patterns on  $M$  that consist of marker-allele assignments and that adhere to parameters  $l$ ,  $g$ , and  $i$ , and
- 5 • where  $e(P)$  is true if and only if chi-squared test on  $P$  using haplotype matrix  $H$  and phenotypes  $Y$  exceeds the given threshold  $x$

## Method

```

39.  $S := \{\}$ 
40. // Number of case and control chromosomes:
10 41.  $pi_A :=$  number of disease-associated chromosomes;
42.  $pi_C :=$  number of control chromosomes;
43.  $pi := pi_A + pi_C$ 
44. // A lower bound for pattern frequency:
45.  $lb := pi_A * pi * x / (pi_C * pi + pi_A * x)$ 
15 46. // Variable for iterating over different patterns:
47.  $P = (p_1, \dots, p_k) := ('*', \dots, '*')$ 
48. for  $i := 1$  to  $k$  {
49. // alleles( $m_i$ ) is the set of alleles of the  $i$ :th marker
50. foreach  $a$  in alleles( $m_i$ ) {
20 51.  $p_i := a$ 
52. // Test pattern  $P$  and all its extensions:
53. checkPatterns( $P, i, i, 0, 0$ )
54. // Reset  $p_i$ :
55.  $p_i := '*'$ 
25 56. }
57. }
58. end

59. // Test haplotype pattern  $P$  and all patterns that can be generated by extending  $P$ 
30 60. // from the right:
61. procedure checkPatterns( $P, start, i, nr\_of\_gaps, gap\_length$ ) {
62. // Output strongly associated patterns
63. if chi-squared( $P, M, H, Y$ )  $\geq x$  and  $p_i \neq '*'$  then insert  $P$  into set  $S$ 
64. // Return if extended patterns would be too long:
35 65. if  $i = k$  or  $i + 1 - start > l$  then return
66. // Return if extended patterns can not be strongly disease-associated:

```

```

67. if frequency of  $P$  in disease-associated chromosomes is less than  $lb$ 
68. then return;
69. // Create and test legal extensions of current pattern  $P$  (3 cases):
70. // 1. Give marker  $i+1$  all possible values:
5   71. foreach  $a$  in alleles( $m_{i+1}$ ) {
    72.  $p_{i+1} := a$ 
    73. checkPatterns ( $P$ ,  $start$ ,  $i+1$ ,  $nr\_of\_gaps$ ,  $0$ )
    74. }
    75. // 2. Introduce a new gap starting at marker  $i+1$ :
10  76. if  $p_i \neq '*'$  and  $nr\_of\_gaps < g$  and  $s \geq 1$  then {
    77.  $p_{i+1} := '*'$ 
    78. checkPatterns ( $P$ ,  $start$ ,  $i+1$ ,  $nr\_of\_gaps+1$ ,  $1$ )
    79. }
    80. // 3. Extend the current gap over marker  $i+1$ :
15  81. if  $p_i = '*'$  and  $gap\_length < s$  then {
    82.  $p_{i+1} := '*'$ 
    83. checkPatterns ( $P$ ,  $start$ ,  $i+1$ ,  $nr\_of\_gaps$ ,  $gap\_length+1$ )
    84. }
    85. // Before returning, reset  $p_{i+1}$ :
20  86.  $p_{i+1} := '*'$ 
    87. return
    88. }

```

7. A method of claim 1, wherein the marker patterns  $P$  are searched by the following algorithm:

- 25 Input
- set  $U$  of marker patterns
  - evaluation function  $e(P)$  for patterns  $P$  in  $U$
  - (generalization) relation  $<$  for patterns in  $U$ , where the function  $e$  and the relation  $<$  are such that if  $e(P)$  is true and  $P' < P$ , then  $e(P')$  is also true
- 30 Output
- set  $S = \{P \text{ in } U \mid e(P) \text{ is true}\}$  of patterns

Definitions

- function  $Lgg: U \rightarrow 2^U$ ,  $Lgg(P) = \{ P' \text{ in } U \mid P > P' \text{ and } P' \neq P \text{ and there is no } P'' \text{ in } U \text{ such that } P \neq P'' \neq P' \text{ and } P > P'' > P' \}$ , the set of least general generalizations of pattern  $P$ .
- 5 • function  $Lss: U \rightarrow 2^U$ ,  $Lss(P) = \{ P' \text{ in } U \mid P < P' \text{ and } P' \neq P \text{ and there is no } P'' \text{ in } U \text{ such that } P \neq P'' \neq P' \text{ and } P < P'' < P' \}$ , the set of least special specializations of pattern  $P$ .

Method

```

89.  $S := \{\}$ 
90.  $Q := \{\}$ 
10 91. // Start with the most general patterns:
92.  $F := \{P \text{ in } U \mid \text{there is no } P' \text{ in } U, P' \neq P, \text{ such that } P' < P\};$ 
93. while  $F \neq \{\}$  {
94.     // Evaluate the candidate patterns:
95.     foreach  $P$  in  $F$  {
15 96.         if  $e(P) = \text{true}$  then insert  $P$  into set  $S$ 
97.         else remove  $P$  from set  $F$ 
98.     }
99.      $Q := Q \text{ union } F$ 
100.    // Generate a new set of candidate patterns:
20 101.     $C := \{\}$ 
102.    foreach  $P$  in  $F$  {
103.         $C := C \text{ union } \{ P' \text{ in } U \mid P' \text{ in } Lss(P) \text{ and for all } P'' \text{ in } Lgg(P):$ 
104.             $P'' \text{ in } Q \}$ 
105.    }
25 106.     $F := C$ 
107. }
108.     end

```

8. A method of claim 1, wherein the marker patterns  $P$  are searched by the following algorithm:
- 30

Input

- set  $U$  of marker patterns
- evaluation function  $e(P)$  for patterns  $P$  in  $U$
- frequency threshold  $x$

## Output

- set  $S = \{P \text{ in } U \mid e(P) \text{ and } ae(P) \text{ is true}\}$  of patterns, where  $ae(P)$  is true if and only if the frequency of pattern  $P$  exceeds a given threshold  $x$

## Definitions

- 5 • function  $Lgg: U \rightarrow 2^U$ ,  $Lgg(P) = \{P' \text{ in } U \mid P \rightarrow P' \text{ and } P' \neq P \text{ and there is no } P'' \text{ in } U \text{ such that } P \neq P'' \neq P' \text{ and } P \rightarrow P'' \rightarrow P'\}$ , the set of least general generalizations of pattern  $P$ .
- 10 • function  $Lss: U \rightarrow 2^U$ ,  $Lss(P) = \{P' \text{ in } U \mid P' \rightarrow P \text{ and } P' \neq P \text{ and there is no } P'' \text{ in } U \text{ such that } P \neq P'' \neq P' \text{ and } P' \rightarrow P'' \rightarrow P\}$ , the set of least special specializations of pattern  $P$ .

## Method

109.  $S := \{\}$
110.  $Q := \{\}$
111. // Start with the most general patterns:
- 15 112.  $F := \{P \text{ in } U \mid \text{there is no } P' \text{ in } U, P' \neq P, \text{ such that } P \rightarrow P'\};$
113. while  $F \neq \{\}$  {
114.     // Evaluate the candidate patterns:
115.     foreach  $P$  in  $F$  {
116.         if  $ae(P) = \text{true}$  then {
- 20 117.             if  $e(P) = \text{true}$  then insert  $P$  into set  $S$
118.         }
119.         else remove  $P$  from set  $F$
120.     }
121.      $Q := Q \text{ union } F$
- 25 122.     // Generate a new set of candidate patterns:
123.      $C := \{\}$
124.     foreach  $P$  in  $F$  {
125.          $C := C \text{ union } \{P' \text{ in } U \mid P' \text{ in } Lss(P) \text{ and for all } P'' \text{ in } Lgg(P'):$
- 30 126.              $P'' \text{ in } Q\}$
127.     }
128.      $F := C$
129. }
130. end
- 35

9. A method of claim 1, wherein

- a) the phenotype being studied is qualitative, and
- b) the pattern evaluation function  $e(P)$  has the form  $e(P) = \text{true if and only if } e'(P) > x$ , where  $e'(P)$  is the (signed) association measure  $\chi^2$  and  $x$  is a user specified minimum value, which is chosen so that the sizes of  $S_i$  are large enough, such as 20, to give statistically sufficiently reliable estimates for the gene locus, and
- c) the score  $s(m_i)$  of marker  $m_i$  is the size of  $S_i$ , also called marker-wise pattern frequency of  $m_i$  and denoted by  $f(m_i)$ .

10. A method of claim 1, wherein

- a) the pattern evaluation function  $e(P)$  has the form  $e(P) = \text{true if and only if } e'(P) > x$ , where  $e'(P)$  is the absolute frequency of pattern  $P$  in the data and  $x$  is a user-specified value, which is chosen so that the sizes of  $S_i$  are large enough, such as 20, to give statistically sufficiently reliable estimates for the gene locus, and,
- b) in order to derive the score  $s(m_i)$ , the p value (statistical significance) of each marker pattern  $P$  in determining the phenotype being studied is evaluated, and
- c) the score  $s(m_i)$  is the distance between the observed p value distribution of patterns in  $S_i$  and the uniform distribution, defined as average of  $(p_i - q_i) \log(p_i / q_i)$  over all  $i = 1..n$ , where  $n$  is the number of haplotype patterns in  $S_i$ ,  $p_i$  is the  $i$ th smallest p value in  $S_i$ , and  $q_i$  is the expectation of the  $i$ th smallest p value, if the p values were randomly drawn from the uniform distribution.

11. A method of claim 10, where the p value is computed using a linear model of form  $Y = \beta_1 X_1 + \dots + \beta_k X_k + \alpha Z + \beta_0$ , where the dependent variable  $Y$  is the phenotype being studied,  $X_1$  through  $X_k$  are covariates, such as environmental factors, and  $Z$  is a dummy variable for the occurrence of the haplotype pattern, and

the coefficients  $\alpha$  and  $\beta_*$  are adjusted for best fit, and then

the significance of  $Z$  as a covariate is assessed using a t test with the null hypothesis " $\alpha = 0$ ".



12. A method of claim 1, wherein each score  $s(m_i)$  is refined by replacing it by the marker-wise p value of the score  $s(m_i)$ , where the statistical significance of  $s(m_i)$  is measured against the null hypotheses that there is no gene effect.
13. A method of claim 12, wherein the marker-wise p values  $p(m_i)$  are determined by randomly permuting phenotypes.
14. A method of claim 1, wherein the area returned from the prediction of the gene location is contiguous or fragmented or a point.
15. A method of claim 1, wherein the location of the gene, predicted as a function of the scores  $s(m_i)$  and based on maximizing or minimizing the score, is predicted to the location of the marker  $m_i$  that maximizes or minimizes the marker score  $s(m_i)$ .
16. A method of claim 1, wherein the location of the gene, predicted as a function of the scores  $s(m_i)$  and based on maximizing or minimizing the score, is predicted to the combination of most probable intervals for containing the trait-susceptibility locus that covers at most the desired proportion  $t$  ( $t \in \{0, 100\%\}$ ) of the original region obtained by taking all such points in the studied chromosomal region whose nearest marker is within the  $k$  best scoring markers, where  $k$  is selected such that the resulting area has length at most  $t$  times the length of the studied region, and where  $k$  is maximal such value.
17. A method of claim 1, wherein the location of the gene, predicted as a function of the scores  $s(m_i)$  and based on maximizing or minimizing the score, is predicted to those points in the studied chromosomal region whose nearest marker scores at least  $y$  or at most  $y$ , where  $y$  is scoring function dependent and is selected so that the probability of the gene being close to the marker is sufficiently large.
18. A method of claim 1, wherein the location of the gene, predicted as a function of the scores  $s(m_i)$  and based on maximizing or minimizing the score, is determined by expert investigation of the marker scores or their visualization.
19. A method of claim 1, wherein several genes are searched for simultaneously by using marker patterns that refer to several potential gene loci at the same time.
20. A computer-readable data storage medium having computer-executable program code stored thereon operative to perform a method of any of preceding claims when executed on a computer.

21. A computer system programmed to perform the method of any of claims 1 to 19.

<p>                     1. 2000-2001                      2. 2001-2002                      3. 2002-2003                      4. 2003-2004                      5. 2004-2005                      6. 2005-2006                      7. 2006-2007                      8. 2007-2008                      9. 2008-2009                      10. 2009-2010                      11. 2010-2011                      12. 2011-2012                      13. 2012-2013                      14. 2013-2014                      15. 2014-2015                      16. 2015-2016                      17. 2016-2017                      18. 2017-2018                      19. 2018-2019                      20. 2019-2020                      21. 2020-2021                      22. 2021-2022                      23. 2022-2023                      24. 2023-2024                      25. 2024-2025                      26. 2025-2026                      27. 2026-2027                      28. 2027-2028                      29. 2028-2029                      30. 2029-2030                      31. 2030-2031                      32. 2031-2032                      33. 2032-2033                      34. 2033-2034                      35. 2034-2035                      36. 2035-2036                      37. 2036-2037                      38. 2037-2038                      39. 2038-2039                      40. 2039-2040                      41. 2040-2041                      42. 2041-2042                      43. 2042-2043                      44. 2043-2044                      45. 2044-2045                      46. 2045-2046                      47. 2046-2047                      48. 2047-2048                      49. 2048-2049                      50. 2049-2050                      51. 2050-2051                      52. 2051-2052                      53. 2052-2053                      54. 2053-2054                      55. 2054-2055                      56. 2055-2056                      57. 2056-2057                      58. 2057-2058                      59. 2058-2059                      60. 2059-2060                      61. 2060-2061                      62. 2061-2062                      63. 2062-2063                      64. 2063-2064                      65. 2064-2065                      66. 2065-2066                      67. 2066-2067                      68. 2067-2068                      69. 2068-2069                      70. 2069-2070                      71. 2070-2071                      72. 2071-2072                      73. 2072-2073                      74. 2073-2074                      75. 2074-2075                      76. 2075-2076                      77. 2076-2077                      78. 2077-2078                      79. 2078-2079                      80. 2079-2080                      81. 2080-2081                      82. 2081-2082                      83. 2082-2083                      84. 2083-2084                      85. 2084-2085                      86. 2085-2086                      87. 2086-2087                      88. 2087-2088                      89. 2088-2089                      90. 2089-2090                      91. 2090-2091                      92. 2091-2092                      93. 2092-2093                      94. 2093-2094                      95. 2094-2095                      96. 2095-2096                      97. 2096-2097                      98. 2097-2098                      99. 2098-2099                      100. 2099-2100                      101. 2100-2101                      102. 2101-2102                      103. 2102-2103                      104. 2103-2104                      105. 2104-2105                      106. 2105-2106                      107. 2106-2107                      108. 2107-2108                      109. 2108-2109                      110. 2109-2110                      111. 2110-2111                      112. 2111-2112                      113. 2112-2113                      114. 2113-2114                      115. 2114-2115                      116. 2115-2116                      117. 2116-2117                      118. 2117-2118                      119. 2118-2119                      120. 2119-2120                      121. 2120-2121                      122. 2121-2122                      123. 2122-2123                      124. 2123-2124                      125. 2124-2125                      126. 2125-2126                      127. 2126-2127                      128. 2127-2128                      129. 2128-2129                      130. 2129-2130                      131. 2130-2131                      132. 2131-2132                      133. 2132-2133                      134. 2133-2134                      135. 2134-2135                      136. 2135-2136                      137. 2136-2137                      138. 2137-2138                      139. 2138-2139                      140. 2139-2140                      141. 2140-2141                      142. 2141-2142                      143. 2142-2143                      144. 2143-2144                      145. 2144-2145                      146. 2145-2146                      147. 2146-2147                      148. 2147-2148                      149. 2148-2149                      150. 2149-2150                      151. 2150-2151                      152. 2151-2152                      153. 2152-2153                      154. 2153-2154                      155. 2154-2155                      156. 2155-2156                      157. 2156-2157                      158. 2157-2158                      159. 2158-2159                      160. 2159-2160                      161. 2160-2161                      162. 2161-2162                      163. 2162-2163                      164. 2163-2164                      165. 2164-2165                      166. 2165-2166                      167. 2166-2167                      168. 2167-2168                      169. 2168-2169                      170. 2169-2170                      171. 2170-2171                      172. 2171-2172                      173. 2172-2173                      174. 2173-2174                      175. 2174-2175                      176. 2175-2176                      177. 2176-2177                      178. 2177-2178                      179. 2178-2179                      180. 2179-2180                      181. 2180-2181                      182. 2181-2182                      183. 2182-2183                      184. 2183-2184                      185. 2184-2185                      186. 2185-2186                      187. 2186-2187                      188. 2187-2188                      189. 2188-2189                      190. 2189-2190                      191. 2190-2191                      192. 2191-2192                      193. 2192-2193                      194. 2193-2194                      195. 2194-2195                      196. 2195-2196                      197. 2196-2197                      198. 2197-2198                      199. 2198-2199                      200. 2199-2200                      201. 2200-2201                      202. 2201-2202                      203. 2202-2203                      204. 2203-2204                      205. 2204-2205                      206. 2205-2206                      207. 2206-2207                      208. 2207-2208                      209. 2208-2209                      210. 2209-2210                      211. 2210-2211                      212. 2211-2212                      213. 2212-2213                      214. 2213-2214                      215. 2214-2215                      216. 2215-2216                      217. 2216-2217                      218. 2217-2218                      219. 2218-2219                      220. 2219-2220                      2</p>	
---	--